Negative observations, induction and the generation of hypotheses

Wouter Voorspoels (wouter.voorspoels@ppw.kuleuven.be)

Chayenne Van Meel (chayenne.vanmeel@student.kuleuven.be)

Gert Storms (gert.storms@ppw.kuleuven.be)
KU Leuven, Department of Experimental Psychology, Tiensestraat, 102
Leuven, 3000, Belgium.

Abstract

In category-based induction tasks, a robust finding is that positive observations raise the judged likelihood of a conclusion and negative observations lower judged likelihood. We present evidence that negative observations can raise the judged likelihood. In particular, we asked participants to judge the likelihood of a conclusion after introducing them to different sets of premises either containing one positive observation or the same positive observation and a negative observation. We found that when the negative observation is dissimilar to the positive observation, willingness to accept a conclusion is raised. Moreover, results from a simultaneous hypothesis generation task suggest that the rise in judged conclusion likelihood is due to a peculiar shift in the hypothesis space of the reasoner, in that the hypothesis with the largest extension, yet still consistent with all premises gains disproportionate popularity when introducing a dissimilar negative observation.

Keywords: induction; non-monotonicity; reasoning; sampling assumptions;

Introduction

People often find themselves in situations that require judgments based on incomplete knowledge, derived from an incomplete set of observations. From experience with traffic lights, we can conclude that red is diagnostic for dangerous situations (positive observations), and we will refrain from crossing the road. However, we have also encountered traffic lights on lonely nights, when there is no traffic. In that situation, red *does not* necessarily indicate danger (negative observation). How do we combine these observations to make a decision about crossing the road? The world is not sufficiently friendly to provide us with an exhaustive set of observations. But we do not want to stay on the same side of the road all our lives. We want to see the other side of the road! Therefore inference to uncertain conclusions, generally referred to as induction, is omnipresent in everyday life and almost equally widely studied in cognitive science (Heit, 2000).

A common paradigm to study induction a is the category-based induction task: Participants are asked to infer the presence of a feature in a conclusion category on the basis of a set of observations. The observations are presented as premises of the argument. For example:

Premise: Tigers have sesamoid bones

Conclusion: Lions have sesamoid bones

A number of regularities have been reported regarding how people respond to such problems, one of which forms the topic of the present paper. Following intuition, but also according to the main theories of inductive reasoning (see, e.g., Heit, 2000), there exists a monotonic relation between the number of observations and the strength of an argument: As more objects displaying the property are observed, a conclusion will be judged more likely (see, e.g., Osherson, Smith, Wilkie, Lpez, & Shafir, 1990). Similarly, as objects are observed that do not have the property, the judged likelihood of a conclusion decreases (e.g., Heussen, Voorspoels, Verheyen, Storms, & Hampton, 2011). We will refer to this general finding as the *monotonicity principle*.

Put differently, the monotonicity principle predicts that a positive observation[1] raises argument strength, and a negative observation[2] lowers argument strength. In case the observation is extremely dissimilar, and thus irrelevant to the conclusion, argument strength remains the same. For example, the likelihood of the conclusion 'Lions have sesamoid bones' is raised by the observation that tigers have sesamoid bones, is lowered by the observation that leopards *do not have* sesamoid bones and remains the same by the observation that *tea cups* do not have sesamoid bones. Additional relevant positive and negative observations will respectively raise and lower the argument strength further (asymptotically, obviously).

Recently we have presented evidence that suggests that negative observations can in some cases increase argument strength, contrary to what the monotonicity principle predict (Heussen et al., 2011). In a forced choice paradigm, participants showed a preference for an argument of the following form as compared to an argument without the second, negative premise:

Mozart's music elicits alpha waves in the brain
Metallica's music *does not elicit* alpha waves in the brain

Bach's music elicits alpha waves in the brain

A potential explanation for the results of (Heussen et al., 2011) is that negative observations point the reasoner to a relevant dimension to base inference upon (Medin, Goldstone, and Gentner (1993)). In the above argument, the negative observation highlights a commonality between *Mozart* and *Bach*, not shared by *Metallica*, i.e., that being classical music

[1] An object displaying the property.
[2] An object that does not display the property

is the crucial feature to base inference upon. As the hypothesis that classical music is the crucial feature gains weight, the argument will be considered stronger. In addition, by explicitly contradicting some of the potential hypotheses (e.g.,all music elicits alpha waves), negative evidence clearly helps in reducing the number of hypotheses. This is expected to raise believe in the hypotheses that remain consistent with the premises after the negative observation.

Outline

The present study aims at further investigating the rise in argument strength following a negative observation. In particular, our aim is twofold. First, we want to replicate the effect in a between-subjects rating task, in which we compare generalization judgments of participants who were not presented with negative observations are compared to judgments by participants who were. This differs considerably from Heussen et al. (2011), where a forced choice paradigm was used. Second, we ask participants to generate hypotheses after introducing the observations. In this way, we can examine how the hypothesis space of people confronted with negative observation changes and how this relates to their generalization judgments.

Following Heussen et al. (2011), we hypothesize that a negative observation will raise the willingness to accept a conclusion whenever it points to a dimension that can be used to make the required generalization. In effect, we expect that the projection of a feature from Mozart's music to Bach's music is facilitated when a negative observation excludes other types of music, and points to *classical music* as the correct extension of the novel feature. Similarly, when projecting a property from Bach's music to Nirvana's music, the projection is expected to be facilitated by a negative observation outside the category that entails both subcategories (music). Adding the premise that the sound of a falling rock *does not* have the property, is thus expected to increase the willingness to project the property to Nirvana's music. Moreover, we expect that the hypothesis space of participants will vary accordingly.

Experiment

Method

Participants Participants were 172 bachelor students psychology who volunteered for course credits.

Materials We used 12 argument topics taken from Heussen et al. (2011) (*music, painters, public figures, types of ships, types of glass, types of displays, water bodies, fruit, water birds, insects, polar animals*). In each topic, a hierrachical structure is present, comprising of a category (e.g., *music*), two subcategories (A: *classical music* and B: *rock music*) and a supordinate category (C: *sound*).

Each of the topics has a base argument built from one premis from subcategory A (e.g., Mozart's music has X). Depending on the condition, negative premises are added from the other subcategory (B), or a different category (C). Thus, either the additional premise contains information regarding a member of subcategory A (e.g., Vivaldi's music), or the premise contains information on a member of subcategory B (e.g., Metallica's music) or the premise presents information from category C (e.g., the sound of a waterfall). Table 1 presents an overview of the premises for the topic *music*.

In Heussen et al. (2011), only one conclusion from subcategory A was used. For the present experiment, we added two conclusion categories to each topic: One from subcategory B (e.g., Nirvana's music), and one from category C (e.g., the sound of a falling rock). The properties that were to be generalized from premises to conclusions, were intuitively realistic characteristics that participants were likely to have very little knowledge about (e.g., contain lycopene; create conversion currents; elicit alpha waves in the brain).

Table 1: an overview of the simtulus material for the topic 'music'. Entries in bold refer to items that are presented in every condition (e.g., the base premise "Mozart's music elicits alpha waves").

Type	Premise	Conclusion
subcategory (A)	**Mozart**, Vivaldi	**Bach**
subcategory (B)	Metallica	**Nirvana**
superordinate (C)	falling rock	**waterfall**

Procedure The experiment had the form of a web-based survey. On each trial, participants were presented with a short scenario describing that specialists in the domain of interest (e.g., neuroscientists) had recently made novel discoveries. This was followed with the premise (or premisses) of an argument. For example, participants were given following premises:

Mozart's music elicits alpha waves in the brain. (1)

After reading the information, participants received two successive tasks. First, in the **hypothesis generation task**, they were asked to come up with a rule underlying the observations, (e.g., "classical music elicits alpha waves in the brain"). They were asked to type their hypothesis in a textbox in one or two sentences. Second, in the **generalization task**, participants were asked to judge how likely the three conclusions associated with the argument were by moving a bar on a continuous scale running from 1 to 100 for each of the conclusions.

For each topic, we constructed six premise sets, varying the type of observations, and the "'sign"' (positive or negative) . For each premise set the exact same three conclusions were judged for likelihood, but the premise set varied across condtions. For the present purpose we will discuss only three conditions that allow crucial comparisons to test for the effect of negative observations. In the base condition, referred to as posA, participants received the base premise, as in (1). In

condition posAnegB, a negative observation from a different subcategory is added to the base premise:

Mozart's music elicits alpha waves in the brain.
Metallica's music ***does not*** *elicit alpha waves in the brain.* (2)

In a third condition, posAnegC a negative observation was added to the base premise, disclosing information on a member of the same superordinate category:

Mozart's music elicits alpha waves in the brain.
The sound of a falling rock ***does not*** *elicit alpha waves in the brain* (3)

In total, 12 x 6 arguments were constructed. The 72 arguments were distributed across 6 lists so that each list contained each of the twelve topics (so participants did not see the same topic twice), and a list contained 2 arguments for every type of premise set (so each participant got two arguments from every condition). The lists were distributed randomly across participants. The order of arguments within a list was random for each participant, as well as the order of the conclusions in each argument. The same two practice items preceded the lists for every participants in order to familiarize the participants with the procedure. These two items were not included in the analyses. The experiment took no longer than 20 minutes

Premise sets posA, posAnegB and posAnegC form the object of the present examination. The structure of premise sets 4 to 6 is listed in Table 2, but will not be discussed in the present paper. As can be seen in table 2, these premise sets do not contain negative observations, except the "completely saturated" premise set 6.

Table 2: Schematic overview of the experiment. '+' refers to a positive observation, '-' to a negative observation. '++' means that two premises from the same subcategory were presented in the corresponding condition. In the present paper we focus on the first three premise sets.

Cond	subcat A	subcat B	cat C	# premises
posA	+			1
posAnegB	+	-		2
posAnegC	+		-	2
4	++			2
5	+	+		2
6	+ +	-		3

Results

Generalization

To recapitulate, participant were shown a set of premises (observations), and asked to judge the likelihood of three conclusions. One conclusions concerned a member of the same subcategory as the base premise (subcategory A), a second conclusion concerned a member of a different subcategory (subcategory B) and a third conclusion a member of the shared superordinate category (category C). In this section, we examine the manner in which these generalization judgments vary as a function of the premise set that is presented, and in particular, whether adding a negative observation to the premise set can raise argument strength. In what follows, it is informative to keep in mind that, according to the monotonicity principle, negative observations are expected to lower the likelihood of a conclusion (or leave it unaltered).

Figure 1 presents the average scores of all three conclusion likelihood judgments, averaged across participants and items, as a function of the premise set. PosA introduces only the base premise, posAnegB adds a negative observation of subcategory B to the base premise and posAnegC adds a negative observation of category C. In the two following sections, statistical analyses are presented to test for the effects of adding negative observations to a premise set.

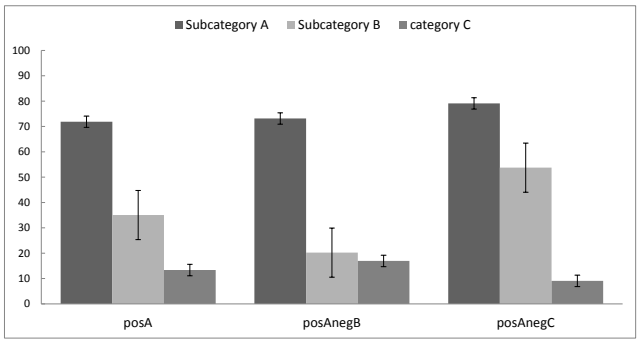

Figure 1: Average judged conclusion likelihood for the three types of conclusions, as a function of the premise set.

Generalizing to the same subcategory For the conclusions that concern a member of subcategory A (e.g., conclusions to *Bach's music* if *Mozart's music* is the base premise), we are interested in two conditions that should lead to a raise in conclusion likelihood as compared to the base argument, despite the negative observation in a premise set. As in Heussen et al. (2011), we expect to observe a difference between premise set posA and posAnegB, in which a negative observation from subcategory B is added to the premise set. Additionally, we hypothesize that adding negative observation from an entirely different category, as in premise set posAnegC, also raises argument strength. In Figure 1 the left bars present the average judged likelihood of the conclusion to a member of subcategory A as a function of the premise set preceding the conclusion, and visual inspection confirms our hypotheses.

We performed a mixed-effects model analyses with two random effects (participants and topics), and two fixed effects (list and premise set), and their interaction[3]. Premise set is a

[3]For the model formulation, we follow (Baayen, Davidson, & Bates, 2008) in their discussion of mixed models for split plot designs. The analyses were carried out in R, using the lme4 package

within subjects factor and list is a between subject variable.

Table 3 gives an overview of the main (fixed) effects of premise set, and can be interpreted as follows: For premise set posA, participants on average judge the likelihood of a conclusion to a member of subcategory A to be 75.14. For premise set posAnegB, in which a negative observation from subcategory B is added, the judged likelihood drops with 2.26 according to the model[4], a change that is not significant. For premise set posAnegC, in which a negative observation from a different category C is added, the judged likelihood is significantly higher 11.12 points ($p = .016$).

Table 3: Effects of premise set on generalizing to a member of subcategory A.

premise set	MCMC estimate	MCMC p-value
posA (base level)	75.14	$< .001$
posAnegB	-2.26	.72
posAnegC	11.12	.016

In sum, we only find partial support for our hypothesis. In particular, premise sets as used in (Heussen et al., 2011), adding a negative observation from a different subcategory, do not lead to a significant rise in argument strength. We do, however, observe a strong rise in argument strength, when a more distant negative observation – from a different category – is added to the premise set.

Generalizing to a different subcategory For the conclusion to a member of subcategory B (e.g., *Nirvana has X*; the base premise is *Mozart*), we hypothesize that a negative observation from a different category (but shared superordinate category, e.g., *the sound of a falling rock*) can raise judged conclusion likelihood. In Figure 1 the average judged likelihood of conclusions to subcategory B for the relevant premise sets is presented in the middle bar of every group, and a rise in mean judged likelihood from premise set posA to premise set posAnegC can be observed. A quantitative test of the difference was performed using mixed model analyses with two random effects (items and participants) and two fixed effects (list and premise set). As in the previous section, this model was preferable to alternative models in terms of AIC and log likelihood deviance.

Table 4 summarizes the effects of premise set. When adding a negative observation from subcategory B, judged likelihood of the conclusion is lowered by 16.55, nearly significantly (p=.08). Note that in this case the premise set contains a negative observation form the same subcategory as the conclusion. More interestingly, when adding a negative observation from a different category (the sound of a falling rock does not have X), judged likelihood increases an impressive 33.61.

Table 4: Effects of premise set on generalizing to a member of subcategory B.

premise set	MCMC estimate	MCMC p-value
posA (base level)	33.93	$< .001$
posAnegB	-16.55	.08
posAnegC	33.61	$< .001$

Conclusions Our analyses of the judged likelihood of the conclusions have revealed convincing evidence that negative observations can raise argument strength in some circumstances. In particular, we found that a negative observation from a seemingly irrelevant category, can substantially raise the judged likelihood of the conclusion to a member of the same subcategory as the base premise as well as to a member of a different subcategory.

Contrary to (Heussen et al., 2011), we do not find support for a rise in judged likelihood of a conclusion to a member of the same subcategory when a negative observation from a different subcategory is introduced. Note that Heussen et al. used a forced choice paradigm, and report effects that, while significant, were very subtle. Perhaps our methodology was not able to identify these effects.

Hypothesis generation

Before making the generalization judgments, participants were asked to generate a hypothesis that they believed explained the observations in the premises. This allows us to peak at the type of hypotheses participants entertained when confronted with negative observations

We differentiated between four types of hypotheses: First, a hypothesis can state that the property is only applicable to the base premise (e.g., "only Bach has X"). Second, a hypothesis can generalize the property to the subcategory of which the base premise is a member (e.g., "all *classical music* has X"), or, the third type, to the entire category ("all music has X") or, in the fourth case, to the entire superordinate category ("all sound has X"). We classified each rule according to its consequential region following this scheme. Hypotheses that did not fit the scheme, for example due to reporting another subcategory, an unspecified subcategory (e.g., "some types of music") or a causal explanation, were coded as "other"[5].

Figure 2 presents the relative frequencies of each type of hypothesis being generated as response to premise sets posA, posAnegB and posAnegC. To quantify and test differences in hypothesis generation between premise sets, we performed logistic regressions with premise set and list as predictors and a binary variable indicating whether the type of hypothesis was generated as dependent variable. The regressions were performed seperately for each type of hypothesis type .

(Bates & Sarkar, 2007).

[4]Note that the model estimate for posAnegB is negative, contrary to what you would expect on the basis of Figure 3 due to addition of list as a factor.

[5]The criterion for classifying was the literal appearance of the intended terms in the rule (and, obviously, in an unambiguous way).

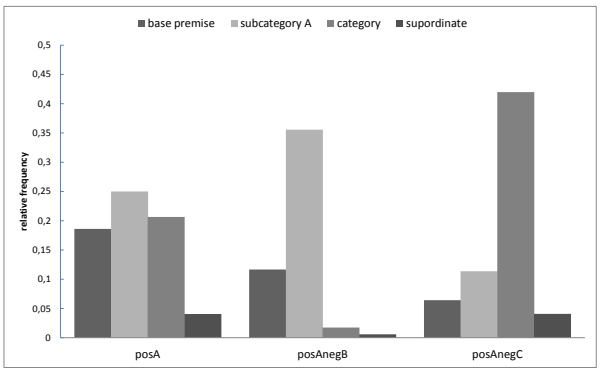

Figure 2: Relative frequency of generating a certain type of hypothesis as a function of the premise set.

Figure 2 suggests three trends are . First, hypotheses that restrict the property to the base premise are significantly less frequent as a negative observation is introduced (with comparison level posA, for posAnegB: $Wald = -2.4, p = .01$; and for posAnegC: $Wald = -4.7, p < .001$). Second, hypotheses that attribute the property to the entire subcategory A are more frequent when a negative observation from subcategory B is introduced (but not significantly, $Wald = .76, p = .45$), and significantly less frequent when a negative observation from category C is introduced ($Wald = -3.74, p < .001$). Third, hypotheses that project the feature to the entire category are less common when introducing a negative observation from a subcategory ($Wald = -6.23, p < .001$), yet more common when introducing a negative observation from a different category ($Wald = 6.02, p < .001$).

Conclusions The hypothesis space in case of arguments with negative observations from subcategory B or category C is substantially altered as compared to the one premise arguments (posA). Moreover, the shifts in generation frequency seem to follow results in the generalization tasks. In particular the increase in subcategory B conclusion likelihood when premise set posAnegC is observed, is clearly associated to an increase in the hypothesis that music is the relevant category.

Not only do some hypotheses become more frequently generated (as can be expected when a number of hypotheses are excluded due to a negative observation), the relative differences between types of hypotheses change considerably across different conditions. In particular, for premise set posAnegB, hypotheses that refer to subcategory A are disproportionately more popular. For premise set posAnegC, hypotheses that refer to the entire category are disproportionately more frequent. Note that for these premise sets, other hypotheses that are also consistent with the observations become less popular. This suggests that more is going on than evenly redistributing the belief of excluded hypotheses across remaining hypotheses. We will come back to this issue in the discussion.

General discussion

We have presented evidence against the universality of the monotonicity principle in inductive reasoning. Negative observations can indeed raise argument strength when they come from a different category than the one shared by the positive observation and the conclusion. Moreover, we found a clear relation with the type of hypotheses that are generated to account for the premise observations. In general, there seems to be a dramatic rise in the weight of the largest hypothesis that is consistent with both positive and negative observations in the premise set. In what follows, we will discuss the relation of these findings to earlier violations of monotonicity in inductive reasoning, and in relation to the sampling assumptions that people have that is, ideas about how the observations are presented to them.

Relation to positive non-monotonicity

For positive observations, a violation of the monotonicity principle has already been documented (Medin, Coley, Storms, & Hayes, 2003), in that under some circumstances positive observations can lower conclusion likelihood. For example, consider following two arguments:

$$\frac{\textit{Brown bears have X}}{\textit{Goats have X}} \tag{4}$$

$$\frac{\textit{Brown bears have X}\quad \textit{Polar bears have X}}{\textit{Goats have X}} \tag{5}$$

Medin et al. (2003) report that participants judge argument (4) stronger than argument (5). According to Medin et al., the addition of the positive observation in (5) reinforces a property that is shared among the premises but is not applicable to the conclusion. Put differently, by adding a positive observation, more weight is given to the hypothesis that the being a bear is crucial for the novel property, and since this property is not shared by the conclusion, it is judged less likely.

The non-monotonicity from adding a negative observation is strikingly symmetric to the non-monotonicity reported by (Medin et al., 2003). Consider following two arguments:

$$\frac{\textit{Mozart's music has X}}{\textit{Metallica's music has X}} \tag{6}$$

$$\frac{\textit{Mozart's music has X}\quad \textit{The sound of a falling rock } \textbf{\textit{does not have}}\ \textit{X}}{\textit{Metallica's music has X}} \tag{7}$$

In the present study, argument (6) was judged stronger by participants. Following our analyses of the hypotheses generated by participants, the addition of the negative observation from outside the music category, drives people to think that "being music" is the most likely hypothesis, rather than, e.g.,

classical music or Mozart's music. By virtue of giving more weight to the music hypothesis, Metallica's music is judged more likely to have X.

In sum, whereas in the positive case, a reasoner's hypothesis "tightens" to a small subcategory (e.g., *bears*) by introducing an observation that is very similar to the base premise (e.g., another type of bear), in the negative case, a reasoner's hypothesis seems to "broaden" to a large category, by introducing an observation that is very dissimilar to the base premise.

Sampling assumptions and non-monotonicity

The question then is how reasoners arrive at weighting exactly these hypotheses more. From a naive probability point of view, excluding certain hypotheses by adding negative observations will automatically lead to a redistribution of the probability mass from the excluded hypotheses to remaining hypotheses. In effect, it makes sense that other hypotheses would indeed become more likely, and as a consequence a particular conclusion could also become more likely. However, it is important to appreciate that, in the naive case, the probability mass would be distributed evenly across the remaining hypotheses, so relative differences between different hypotheses remain. This does not seem to hold in the present results. Indeed, when a negative observation from a different category is observed, as in (6), participants generated the category hypothesis (all music has X) disproportionally more frequently. While consistent with the observations in (6), the subcategory hypothesis (all *classical music* has X) and the base premise hypothesis (*Mozart* has X) experience a substantial drop in generation frequency, contrary to what one would expect on the basis of naive probability theory.

Interestingly, the manner in which Bayesian models of induction (e.g., Tenenbaum & Griffiths, 2001) cope with the positive non-monotonicty effect, is by reweighting the remaining hypotheses when an observation is made. More precisely, depending on assumptions on how the particular observation is sampled from the environment, a Bayesian model would predict that reasoners give more weight as a consistent hypothesis is smaller (e.g., Navarro, Dry, & Lee, 2011; Tenenbaum & Griffiths, 2001.

While technical adjustments to Bayesian inference in nature[6], sampling assumptions also represent a psychological reality and implement what the reasoners' assumptions are on how the observations are presented to him . For example, if a reasoner assumes that the observations in (5) are sampled from the correct hypothesis (for example, because he or she thinks the experimenter intentionally is trying to reveal the correct hypothesis), it is rational to attribute more believe to the hypothesis that it is about bears. Yet, if the reasoner believes the observations are made randomly in the world, and he or she might as well have observed a refridgerator (presumably not having the property in that case) instead of a polar bear, the hypothesis that it is about bears does not gain relative importance (for a more elaborate discussion, see Navarro et al., 2011). Indeed, it would be bad luck on part of the reasoner that he or she did not encounter a more informative observation. The non-monotonicity effects, both positive and negative, suggest that reasoners do not share that assumption.

While the specific implementation of sampling assumptions discussed does not yet apply to negative evidence, a similar reweighting mechanism might be at work when a reasoner is presented with negative evidence. Perhaps reasoners assume that a negative observation is intentionally presented, in such a way that it does not only exclude inconsistent hypotheses, but is informative as to which hypothesis is the correct one.

[6]More specifically, in a weak sampling scheme, hypotheses are not reweighted. In a strong sampling scheme hypotheses with a smaller extension are given more weight

Acknowledgments

Wouter Voorspoels is a postdoctoral fellow at the Research Association - Flanders. We thank Steven verheyen for useful comments.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.

Bates, D. M., & Sarkar, D. (2007). lme4: Linear mixed-effects models using s4 classes, r package version 0.99875-6. [Computer software manual].

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, *7*, 569–592.

Heussen, D., Voorspoels, W., Verheyen, S., Storms, G., & Hampton, J. A. (2011). Raising argument strength using negative evidence: A constraint on models of induction. *Journal of Memory /& Cognition*, *39*, 1496-1507.

Medin, D., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*(3), 517 - 532.

Medin, D., Goldstone, R., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254–278.

Navarro, D. J., Dry, M. J., & Lee, M. D. (2011). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*, 187-223.

Osherson, D., Smith, E., Wilkie, O., Lpez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185–200.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain sciences*, *24*, 762-778.